DeepQuarantine for Suspicious Mail

Никита Беньковия
Kaspersky



Agenda

Спам в мировом трафике

Типы антиспам-решений

DeepQuarantine

Эксперименты атак на DeepQuarantine

Q&A-секция

Agenda

Спам в мировом трафике

Типы антиспам-решений

DeepQuarantine

Эксперименты атак на DeepQuarantine

Q&A-секция

4 млрд

Количество пользователей электронной почты в 2020 году

https://www.statista.com/statistics/255080/number-of-e-mail-users-worldwide/

Спам в мировом трафике в 2020

50.37%
Средний процент спама в мировом трафике

184 млн
Кол-во вредоносных файлов, которые были обнаружены Kaspersky

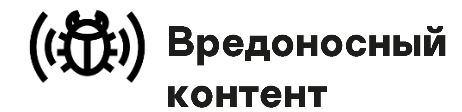
435 млн
Кол-во фишинг-атак, которые были обнаружены Kaspersky



Снижение производительности



Потеря свободного места



https://securelist.com/spam-and-phishing-in-2020/100512/

Agenda

Спам в мировом трафике

Типы антиспам-решений

DeepQuarantine

Эксперименты атак на DeepQuarantine

Q&A-секция

Антиспам-решения в разрезе используемых типов данных

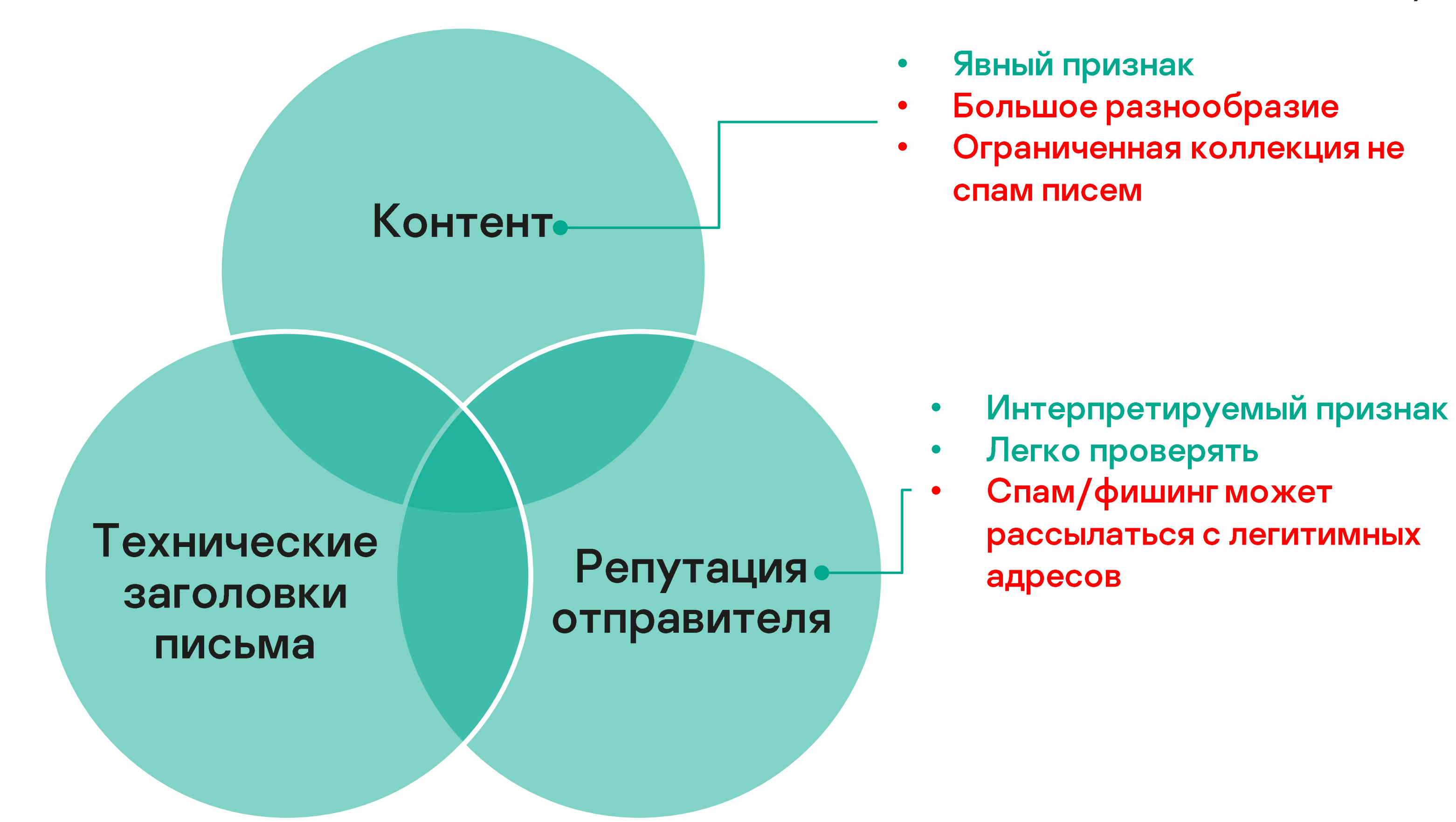


Антиспам-решения в разрезе используемых типов данных

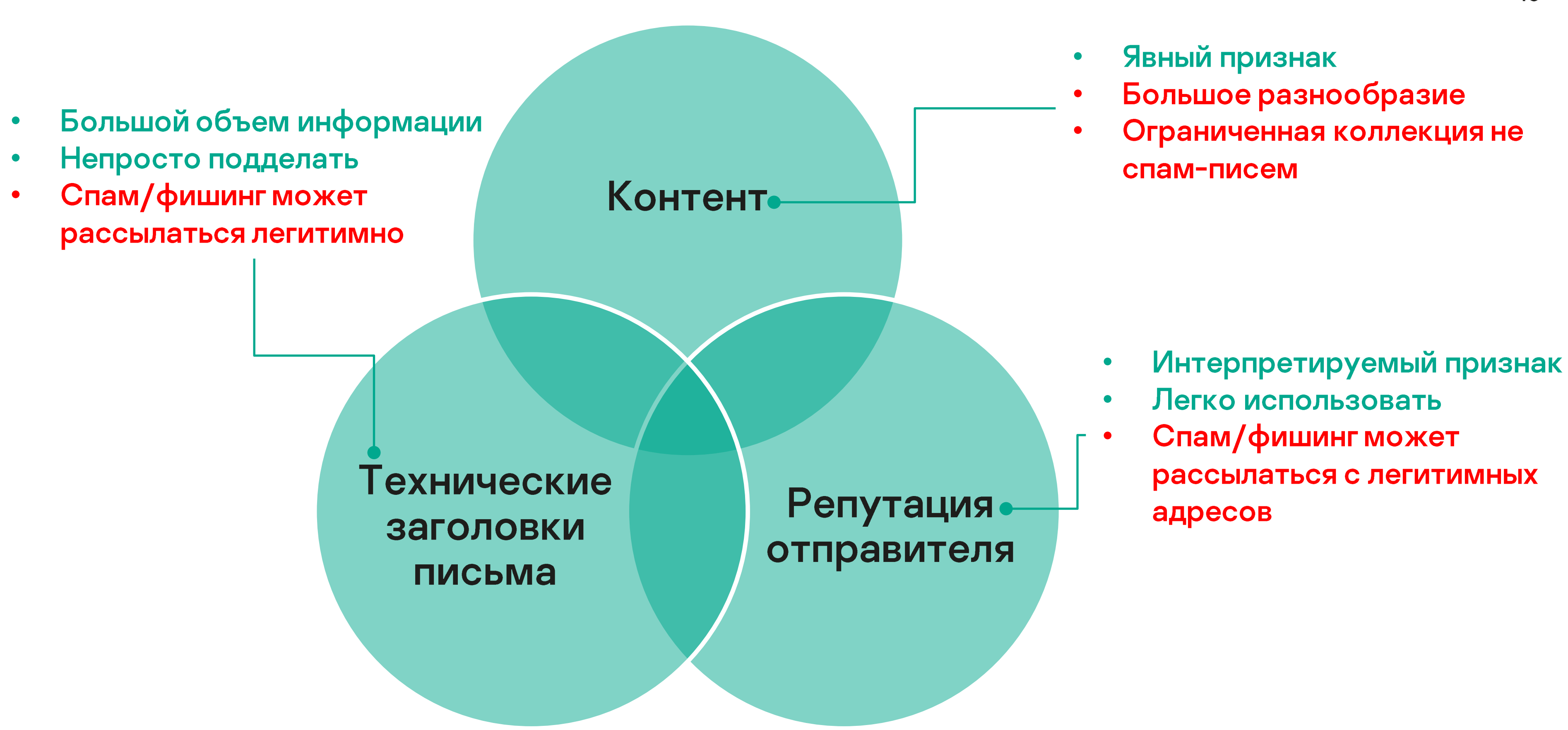


- Явный признак
- Большое разнообразие
- Ограниченная коллекция не спам писем

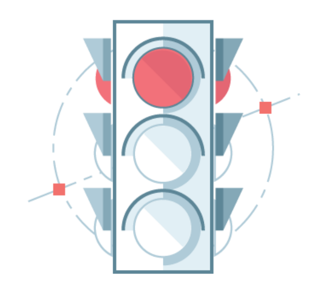
Антиспам-решения в разрезе используемых типов данных



Антиспам-решения в разрезе используемых типов данных



Anti-spam trade-off



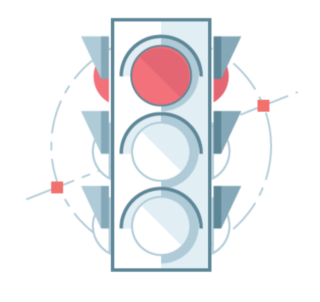




Recall

Precision

Anti-spam trade-off



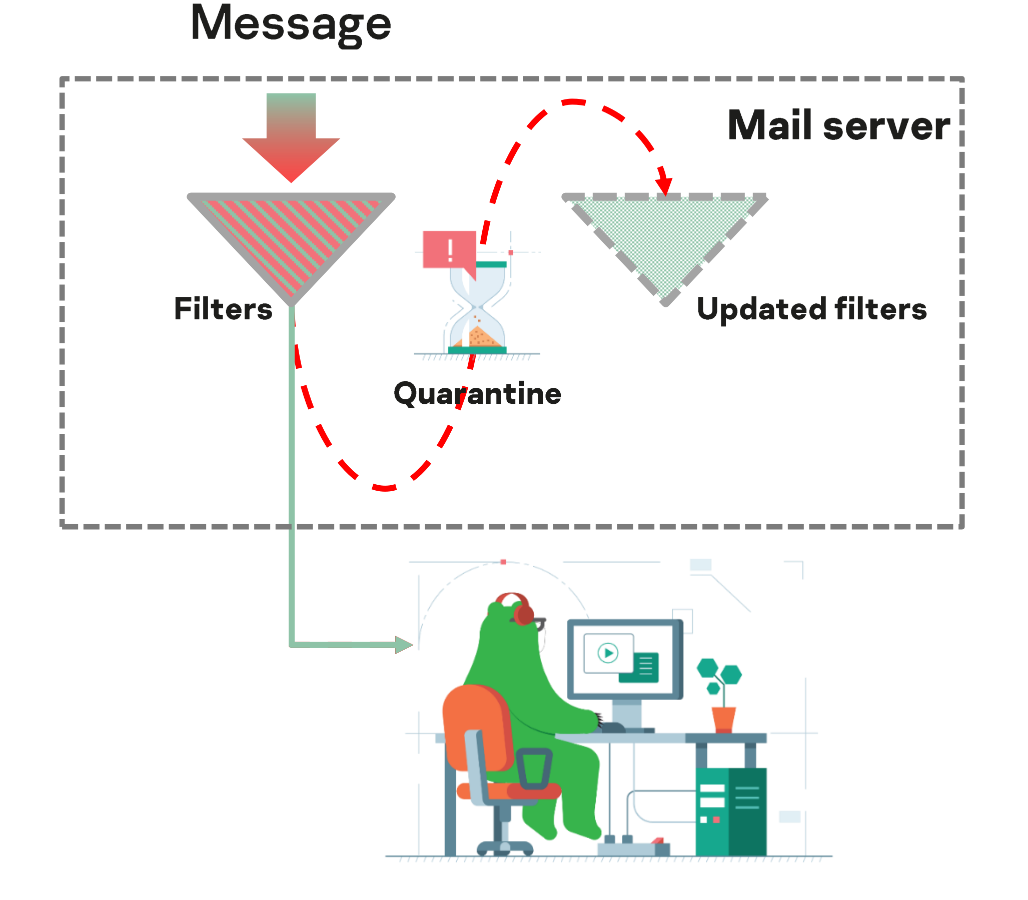




Recall

Precision

На карантин попадают подозрительные письма, чтобы пройти повторную проверку



Agenda

DeepQuarantine

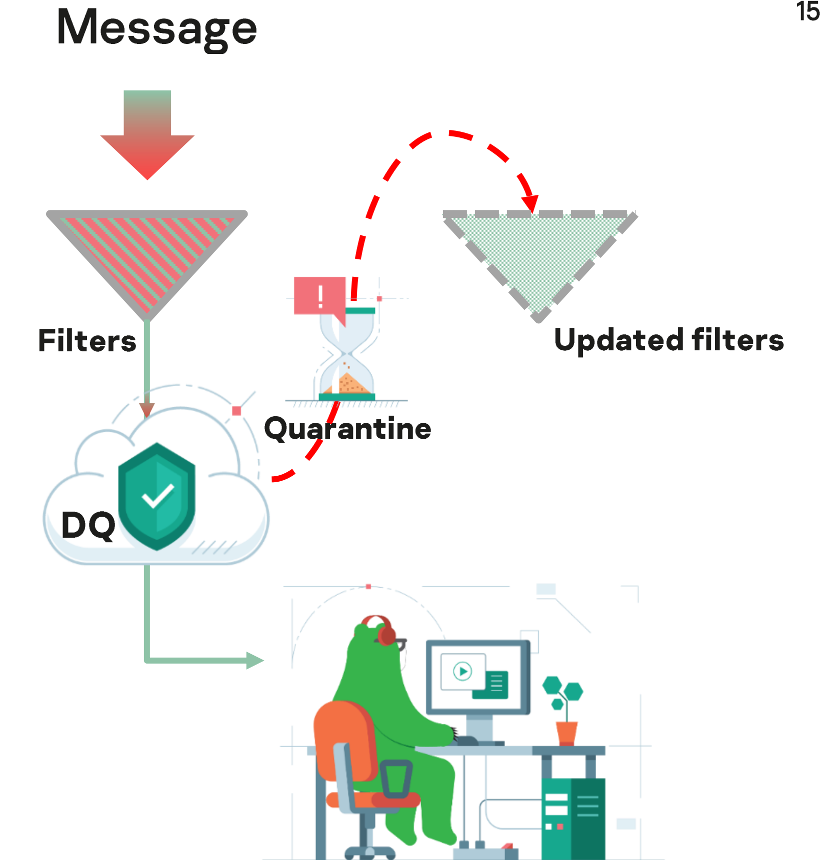
Облачная технология для обнаружения подозрительных писем

Простая интеграция с продуктом

Высокие вычислительные ресурсы

Простая схема обновления модели

Все письма остаются у клиента



https://arxiv.org/abs/2001.04168

Спамеры зачастую используют собственный **Mail User Agent** (MUA) для распространения спама

Для успешной отправки письма они должны корректно заполнить **заголовки MIME**

Let’s check it out!

Subject: I want to steal your personal data!
From: sender@foo.com
To: me@test.com
Date: Mon, 23 Sep 2019 17:00:14 +0300
Message-Id: <h5ced853647a4fd3689a26db412fa4c1@foo.com>
Content-type: multipart/mixed; boundary="===============6411753208318154896=="
X-Mailer: Microsoft Windows Live Mail 14.0.8117.416

Extract features from message

Message-Id
Уникальный идентификатор письма.

Sequence of headers
Последовательность заголовков MIME.

X-mailer
Имя почтового агента.

Признаки

Subject: I want to steal your personal data!

From: sender@foo.com

To: me@test.com

Date: Mon, 23 Sep 2019 17:00:14 +0300

Message-Id: <h5ced853647a4fd3689a26db412fa4c1@foo.com>

Content-type: multipart/mixed; boundary="===============6411753208318154896=="

X-Mailer: Microsoft Windows Live Mail 14.0.8117.416

Extract features from message

Message-Id
Уникальный идентификатор письма.

Sequence of headers
Последовательность заголовков MIME.

X-mailer
Имя почтового агента.

Признаки

Subject: I want to steal your personal data!
From: sender@foo.com
To: me@test.com
Date: Mon, 23 Sep 2019 17:00:14 +0300
Message-Id: <h5ced853647da4fd3689a26db412fa4c1@foo.com>
Content-type: multipart/mixed; boundary="===============6411753208318154896=="
X-Mailer: Microsoft Windows Live Mail 14.0.8117.416

Extract features from message

Message-Id
Уникальный идентификатор письма.

Sequence of headers
Последовательность заголовков MIME.

X-mailer
Имя почтового агента.

Признаки

Subject: I want to steal your personal data!
From: sender@foo.com
To: me@test.com
Date: Mon, 23 Sep 2019 17:00:14 +0300
Message-Id: <h5ced853647a4fd3689a26db412fa4c1@foo.com>
Content-type: multipart/mixed; boundary="===============6411753208318154896=="
X-Mailer: Microsoft Windows Live Mail 14.0.8117.416

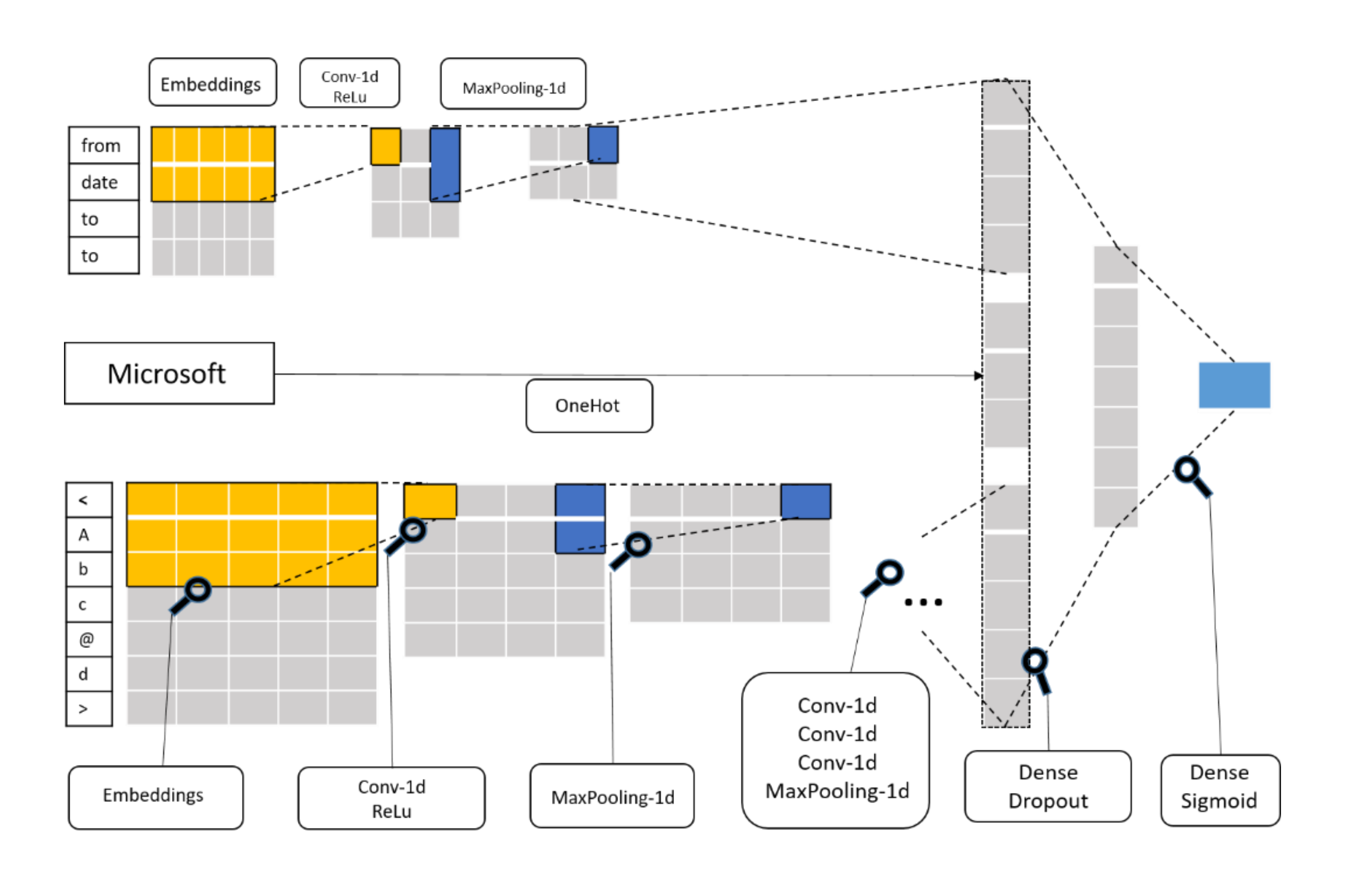
Extract features from message

Message-Id
Уникальный идентификатор письма.

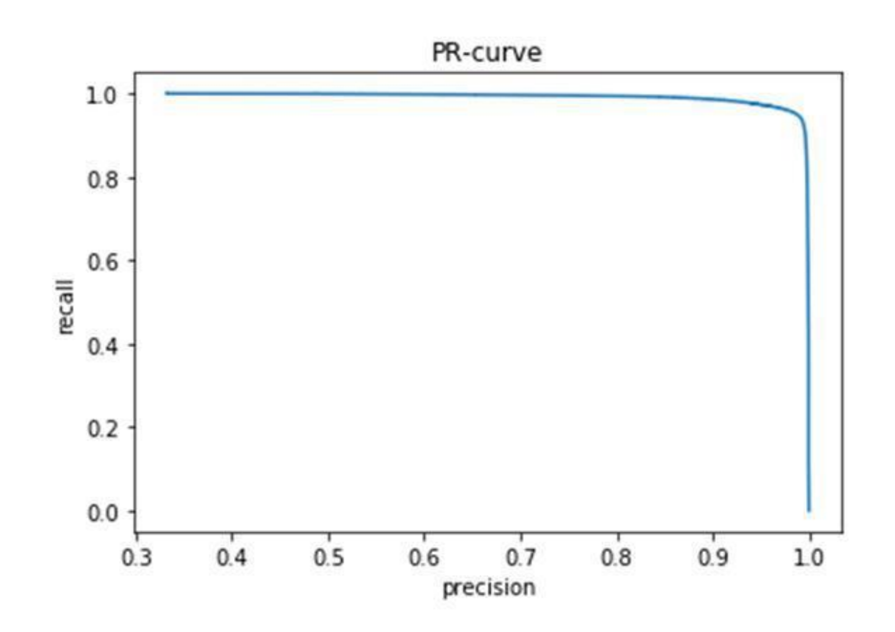
Sequence of headers
Последовательность заголовков MIME.

X-mailer
Имя почтового агента.

Архитектура сети



https://arxiv.org/abs/2001.04168



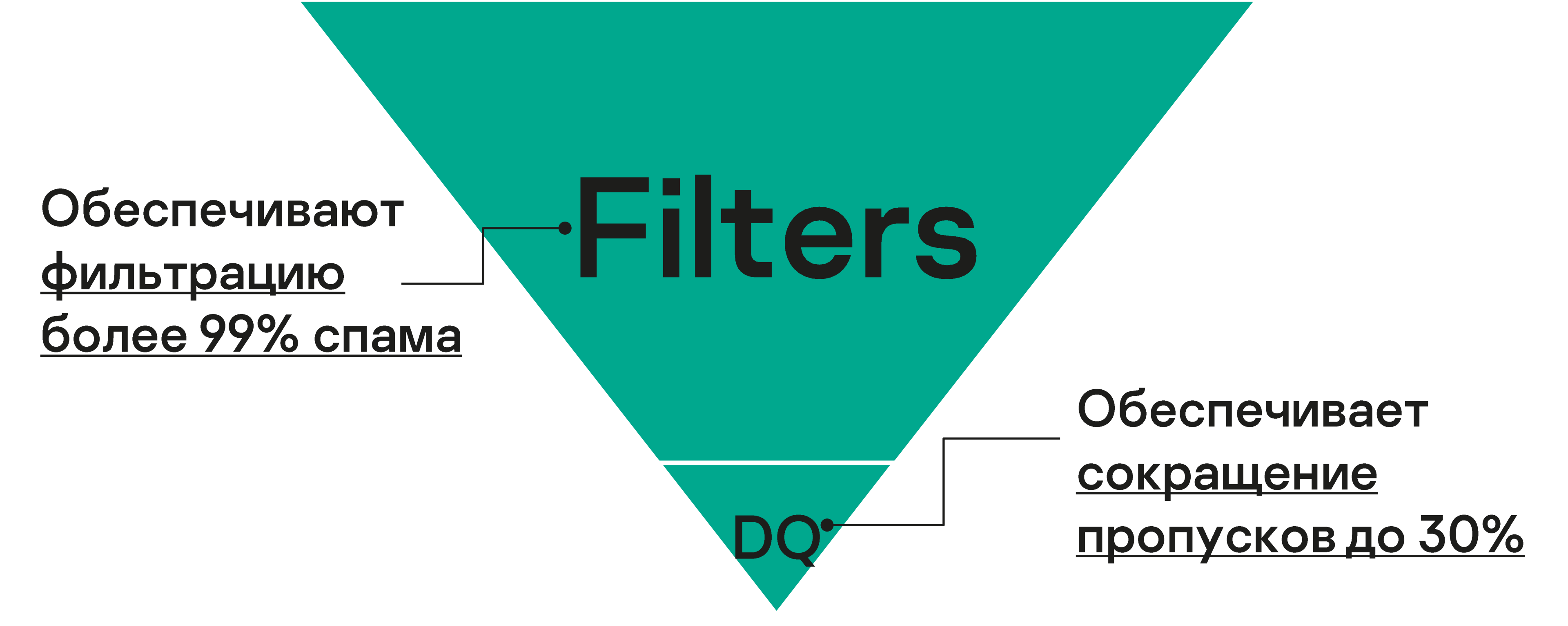
Training

- 120 млн объектов
- 40% спама
- 9 эпох
- SGD с моментом, равным 0.9
- Уменьшаем шаг обучения каждые 3 эпохи

Test

- 40 млн объектов
- 40% спама

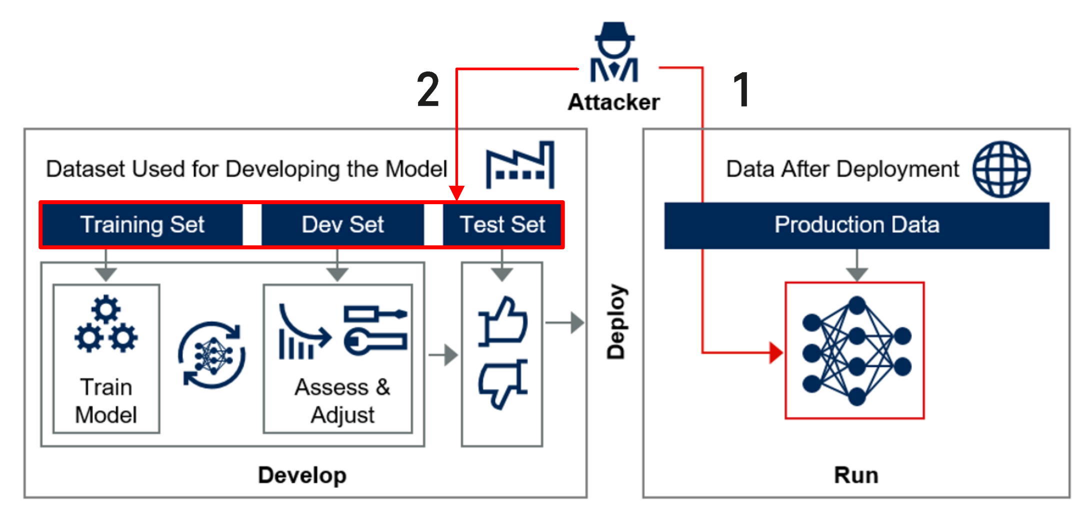
https://arxiv.org/abs/2001.04168


Filters
DQ
Обеспечивают фильтрацию более 99% спама
Обеспечивает сокращение пропусков до 30%

Agenda

Как можно атаковать DeepQuarantine?

Типы атак на нейронные сети



(1) **Adversarial inputs**

Генерация входа с целью уклонения от вердиктов модели

(2) **Data Poisoning**

Влияние на обучающую выборку с целью получения смещенной модели

Использование Data Poisoning для Google Spam Classifier

Model skewing

Загрязнение датасета данными определенного класса с целью смещения decision boundary



Gmail-трафик спам- и не спам-писем. Выделено как минимум четыре масштабные попытки исказить классификатор с помощью отправки большого кол-ва спам-писем как не спам.

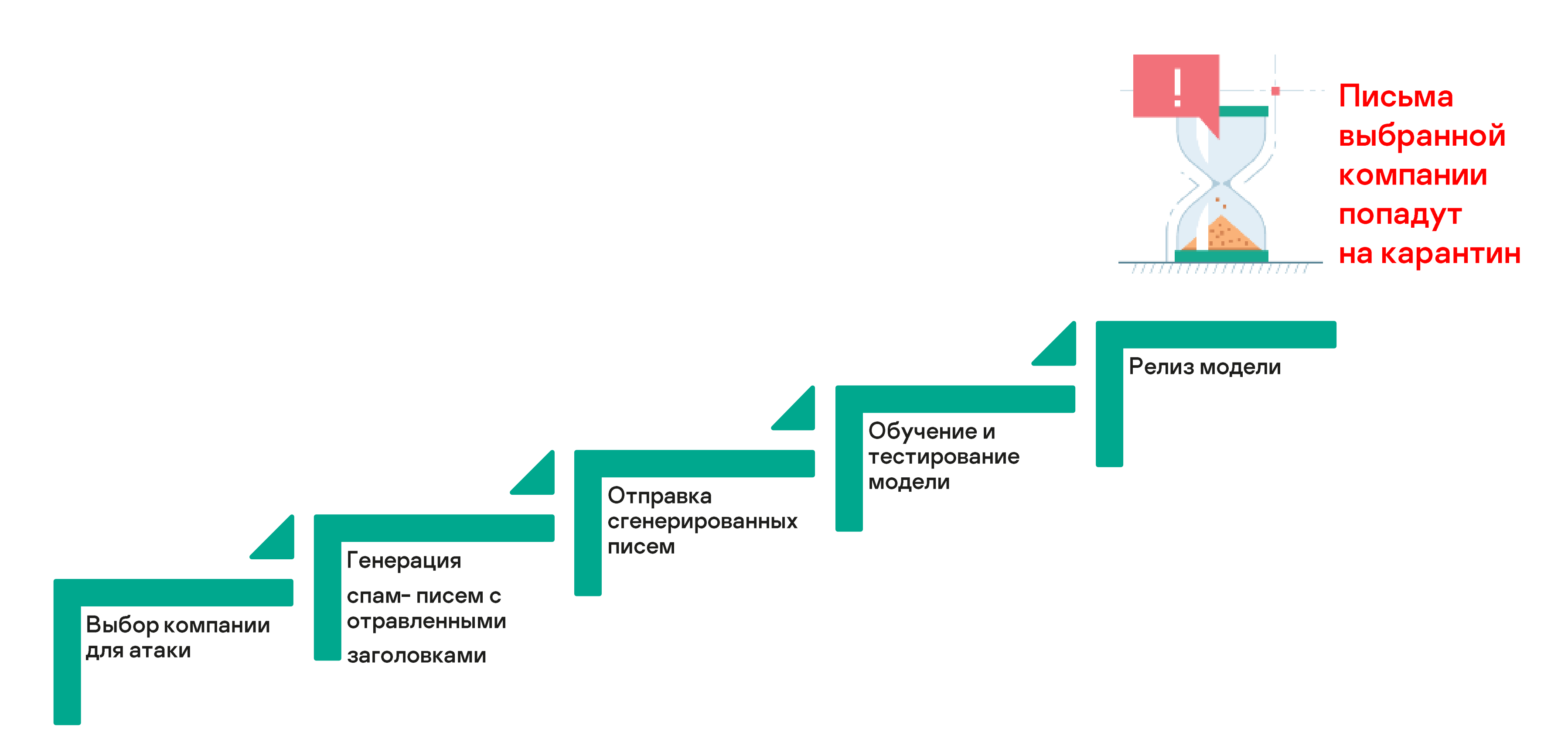
Митигирование рисков

Use sensible data sampling
Необходимо контролировать, чтобы небольшая группа пользователей/IPs не составляла значимую часть обучающей выборки

Compare newly trained model to the previous one
Необходимо сравнивать старую и новую модель с помощью dark launch, A/B или backtesting

Build a golden dataset
Необходимо создать датасет с различными классами, на котором классификатор должен быть предельно точен

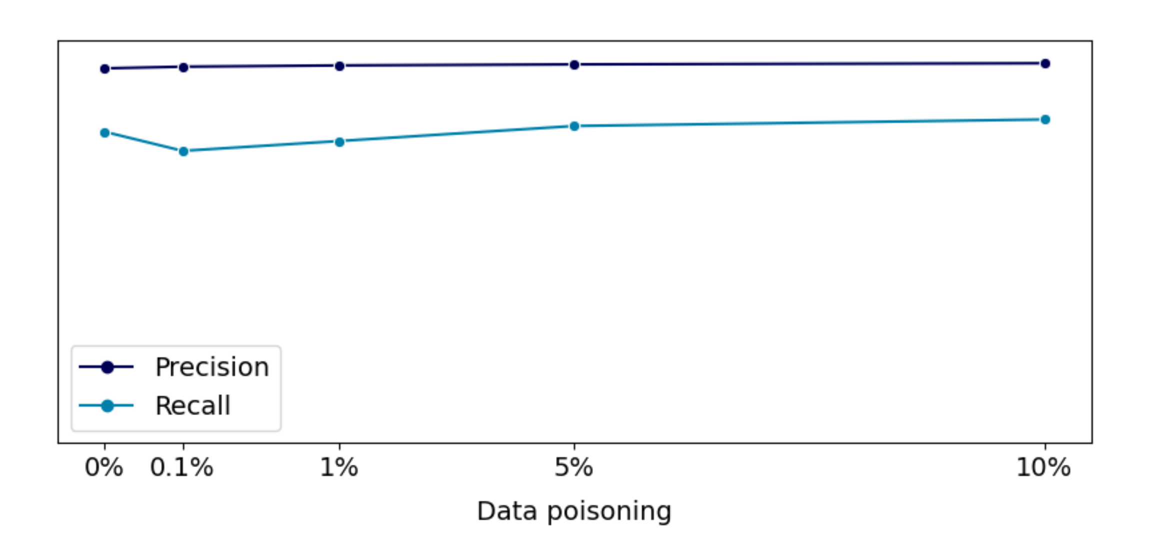
Шаги атаки на DeepQuarantine



Эксперимент 1. Model skewing

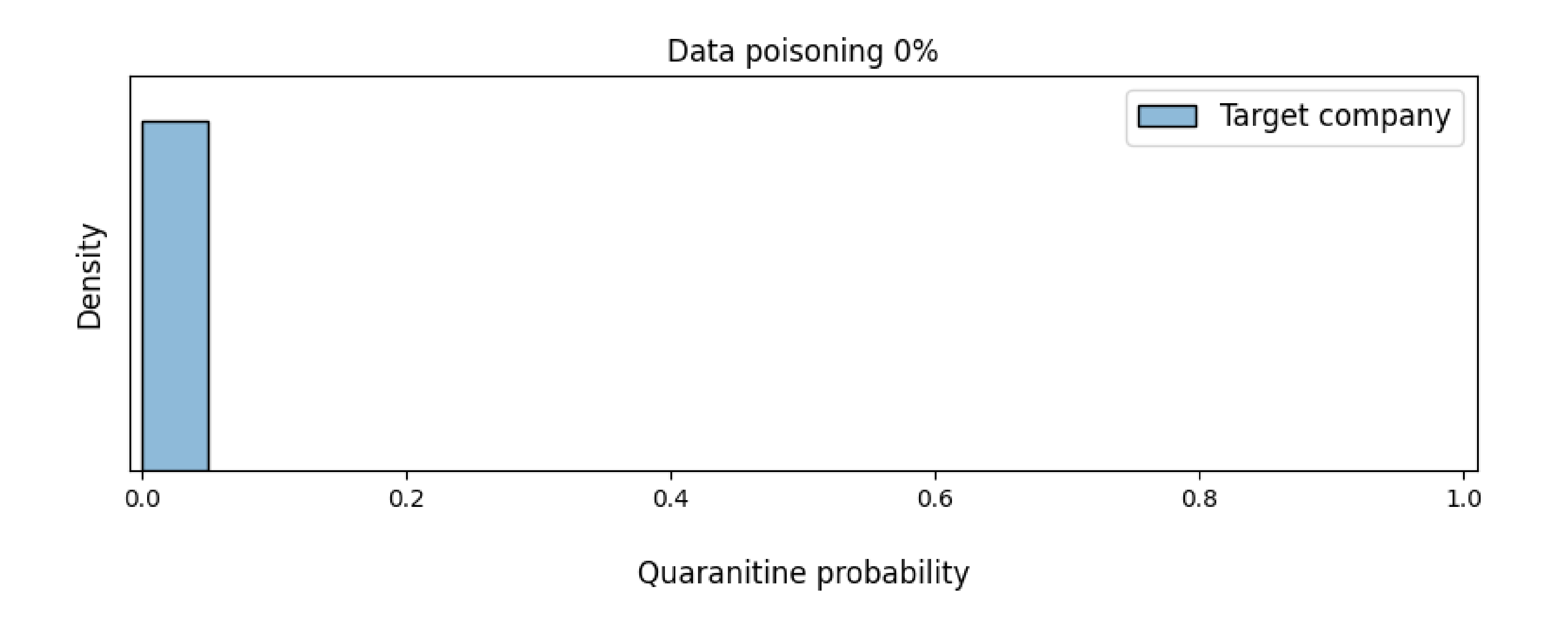
Message-ID: *<........@targeted-company.com>*
Sequence of headers: *const*
X-mailer: *const*

Эксперимент 1. Model skewing

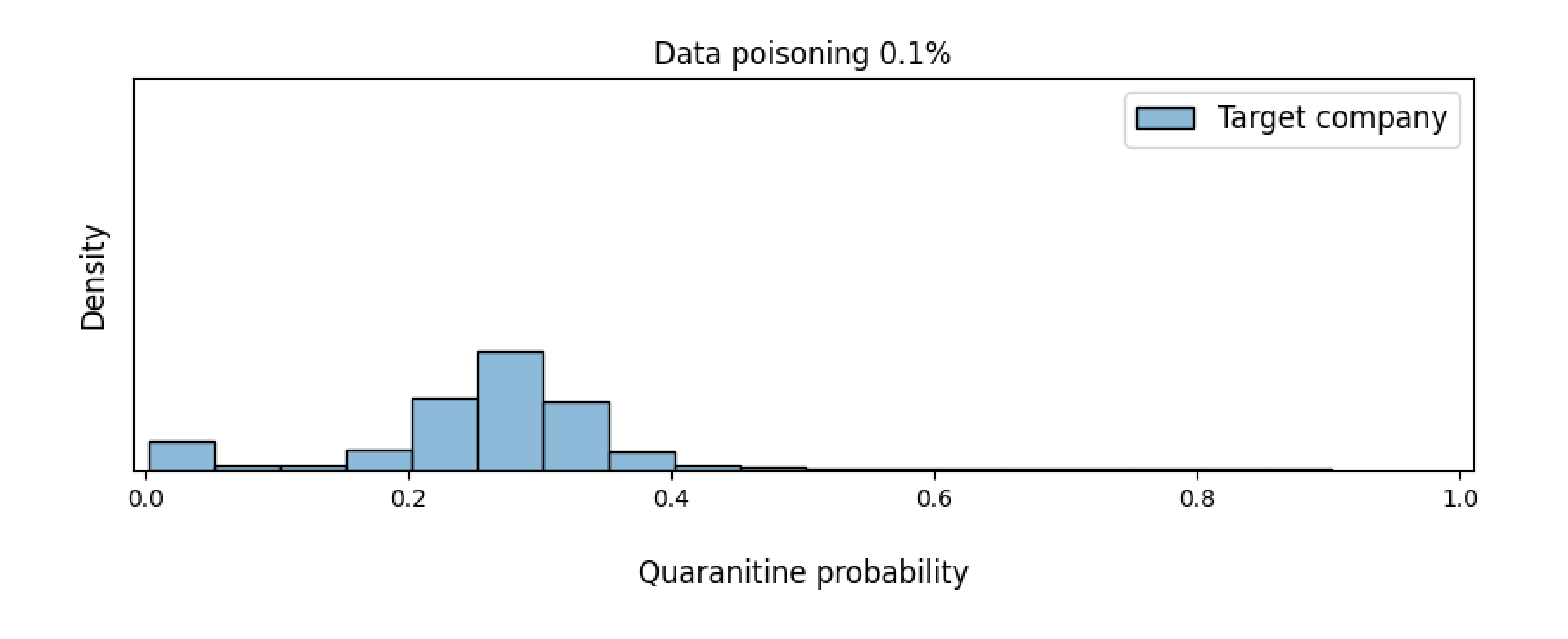


Метрики на валидации не понижаются в зависимости от процента отравления

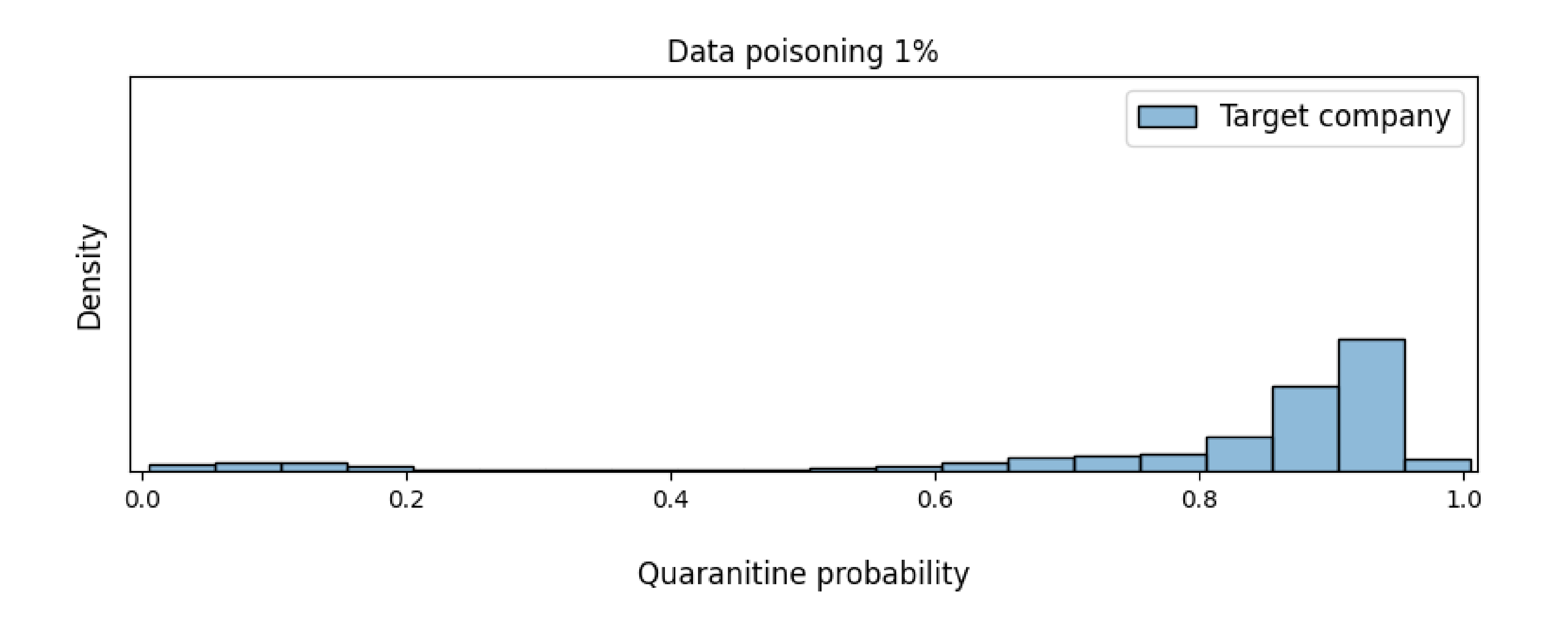
Эксперимент 1. Model skewing



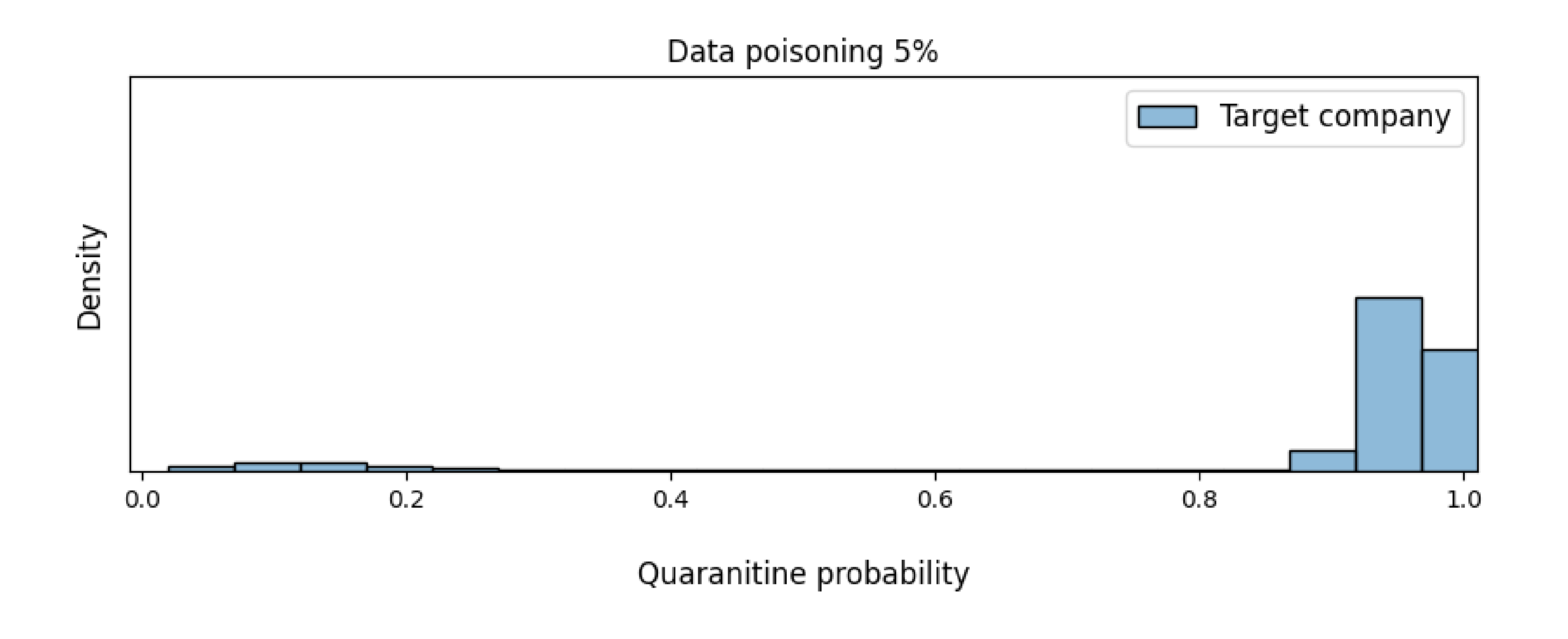
Эксперимент 1. Model skewing



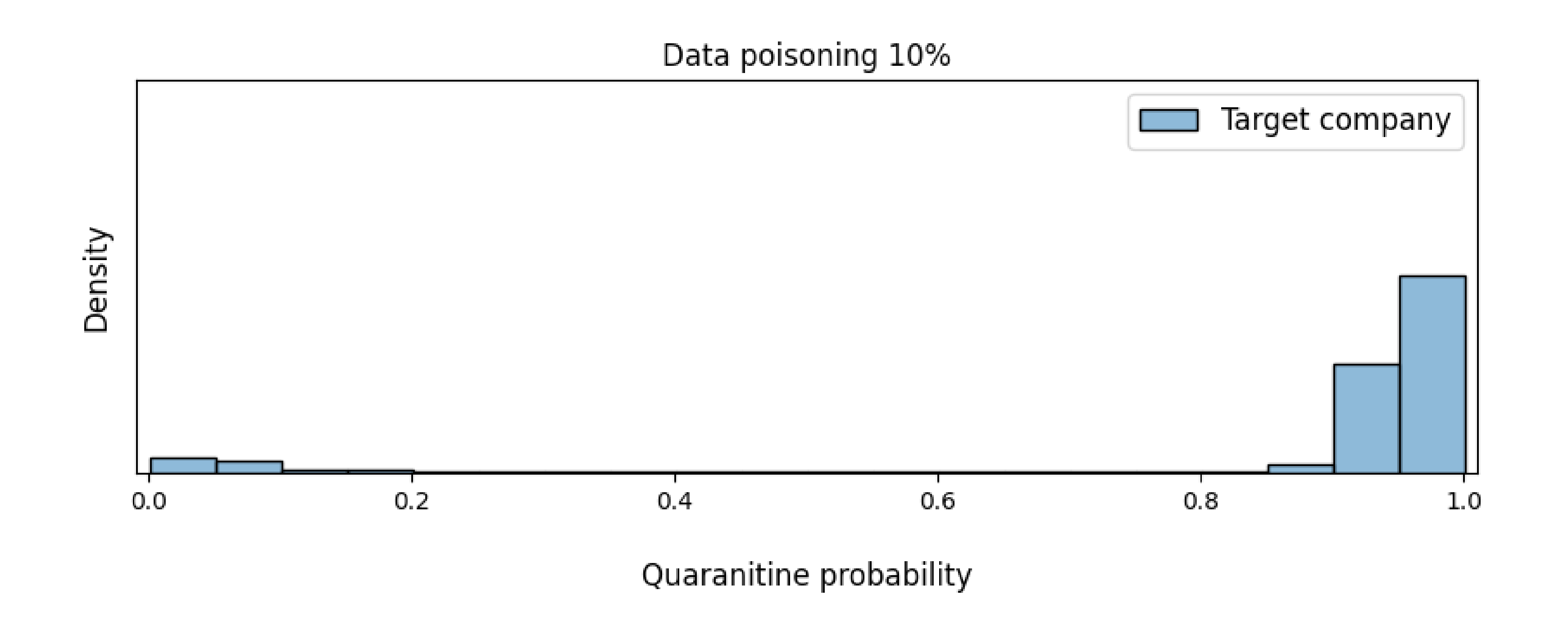
Эксперимент 1. Model skewing



Эксперимент 1. Model skewing



Эксперимент 1. Model skewing

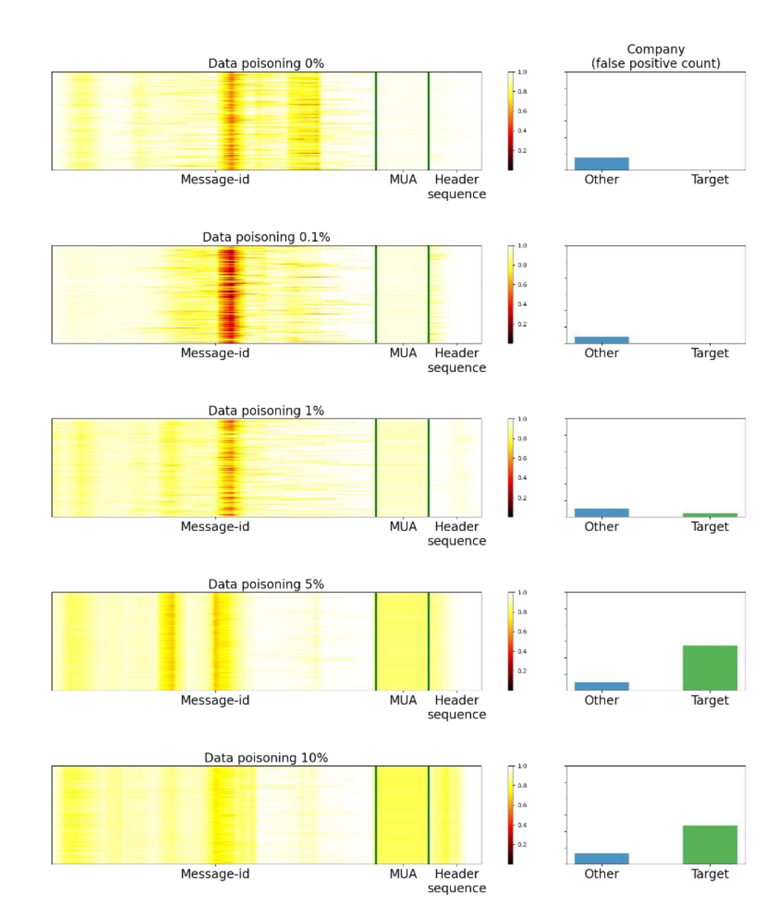


Эксперимент 1. Model skewing

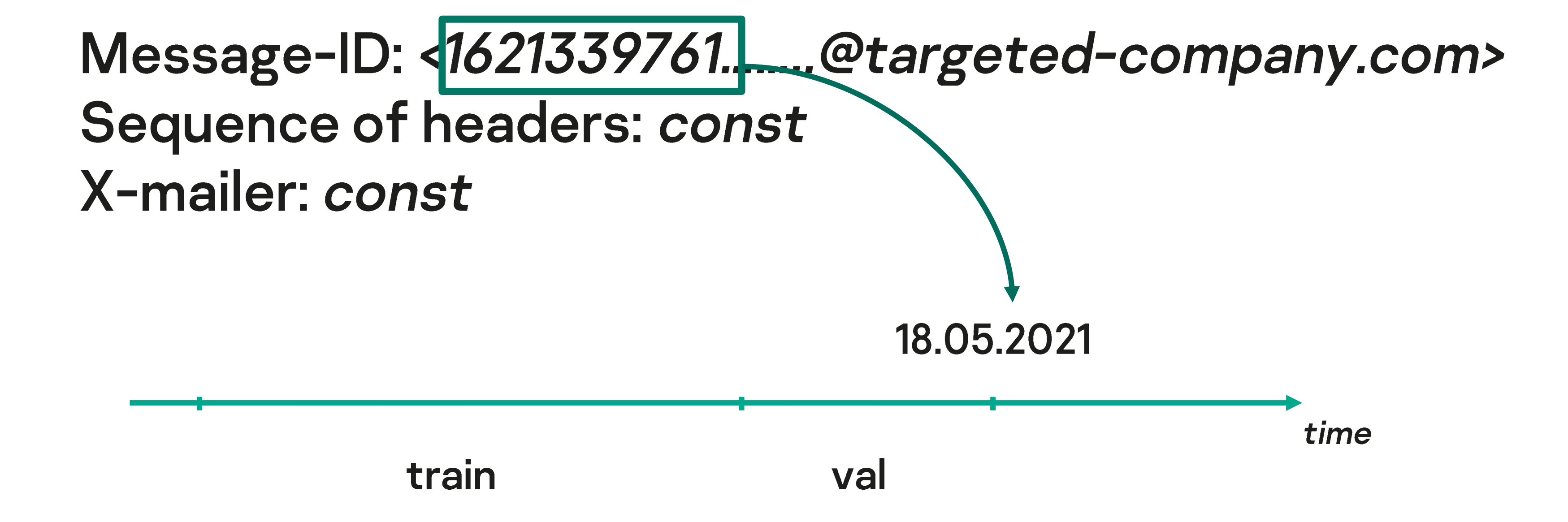
Saliency map

Изображение, которое используется в области компьютерного зрения для определения важности каждого пикселя

Будем обнулять эмбединги на FP-объектах и смотреть, как сильно меняется предсказание модели

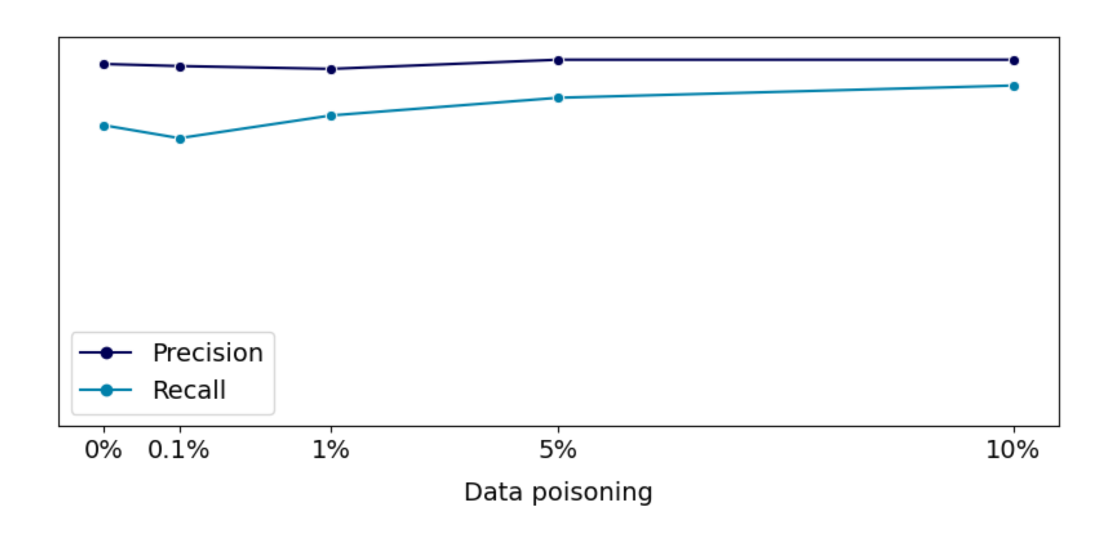


Эксперимент 2. Лик в виде timestamp



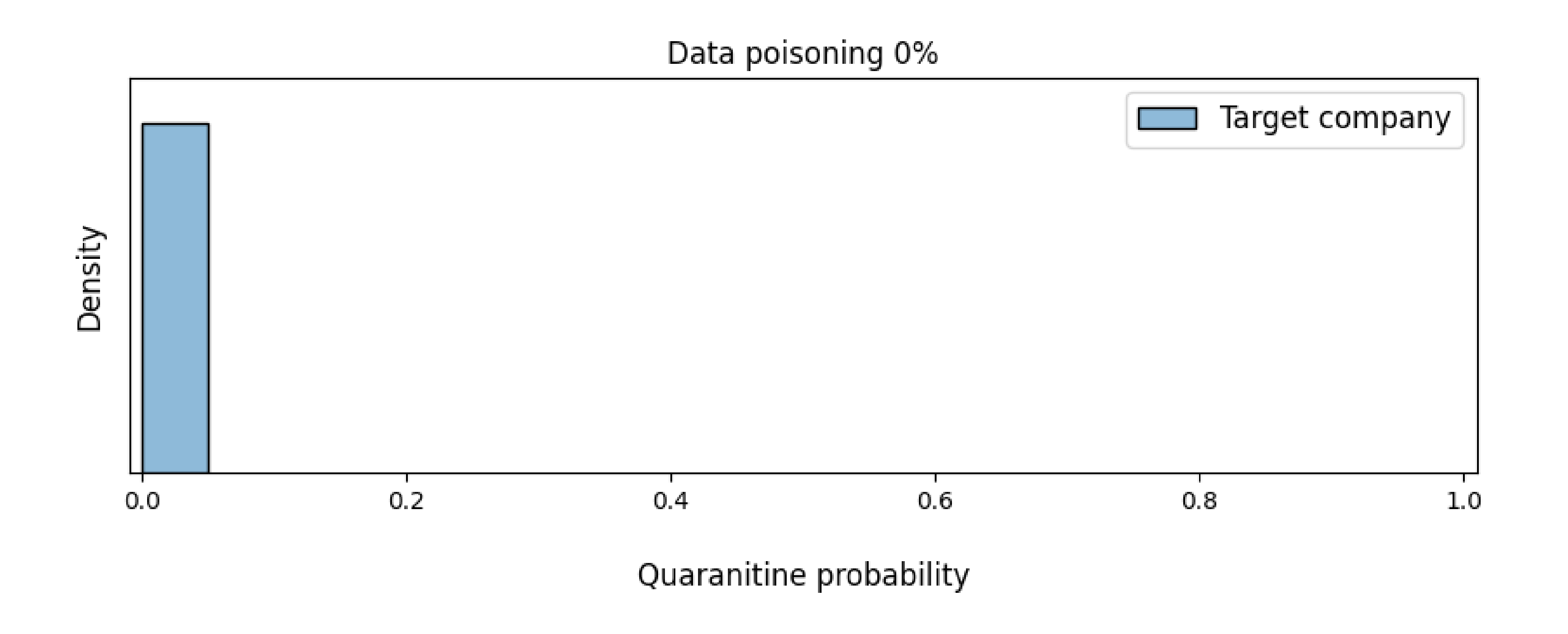
Эксперимент 2. Лик в виде timestamp

Метрики качества на валидации при разном проценте отравления данных

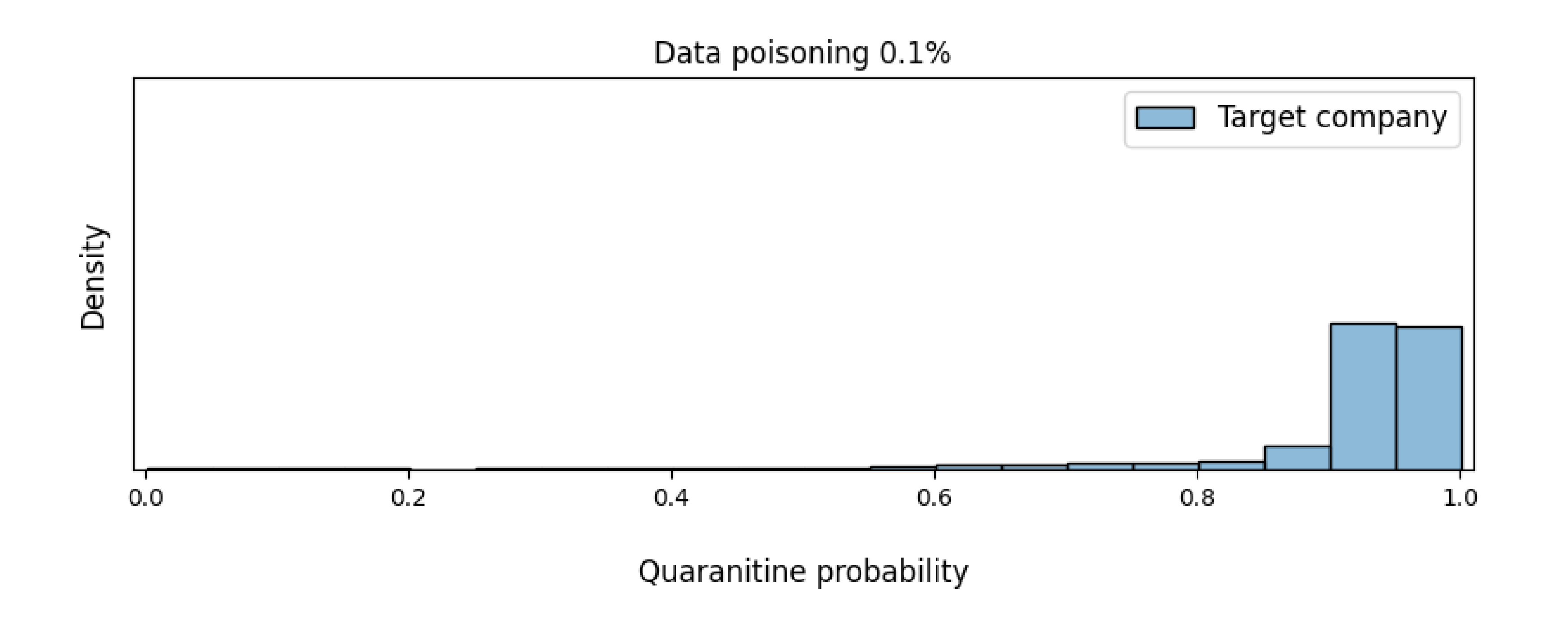


Метрики на валидации не понижаются в зависимости от процента отравления

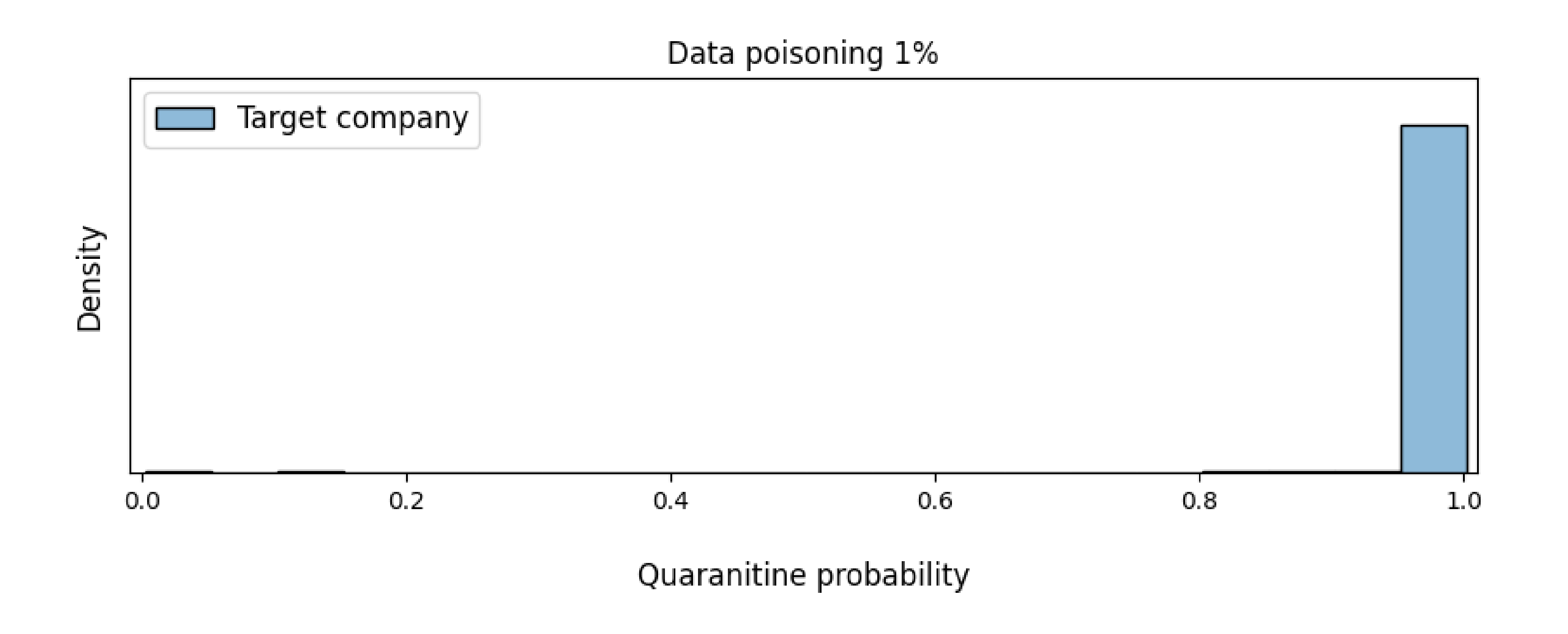
Эксперимент 2. Лик в виде timestamp



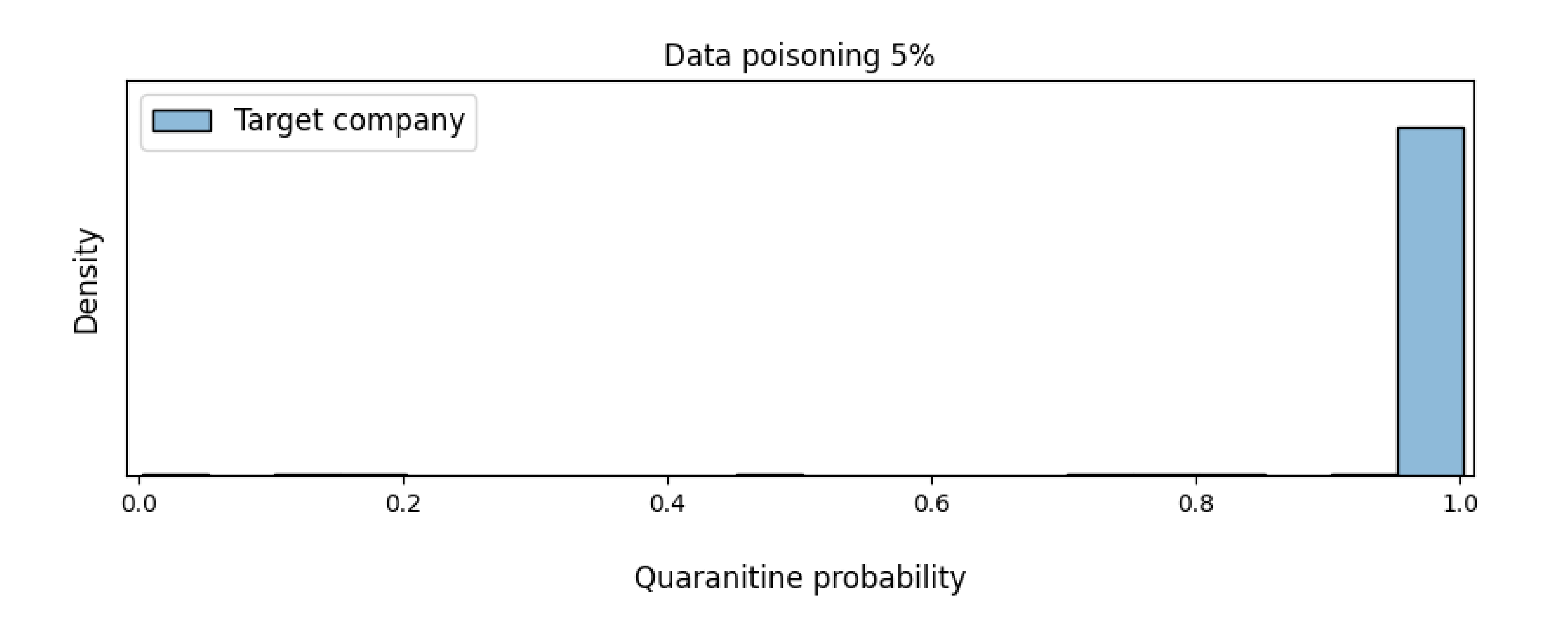
Эксперимент 2. Лик в виде timestamp



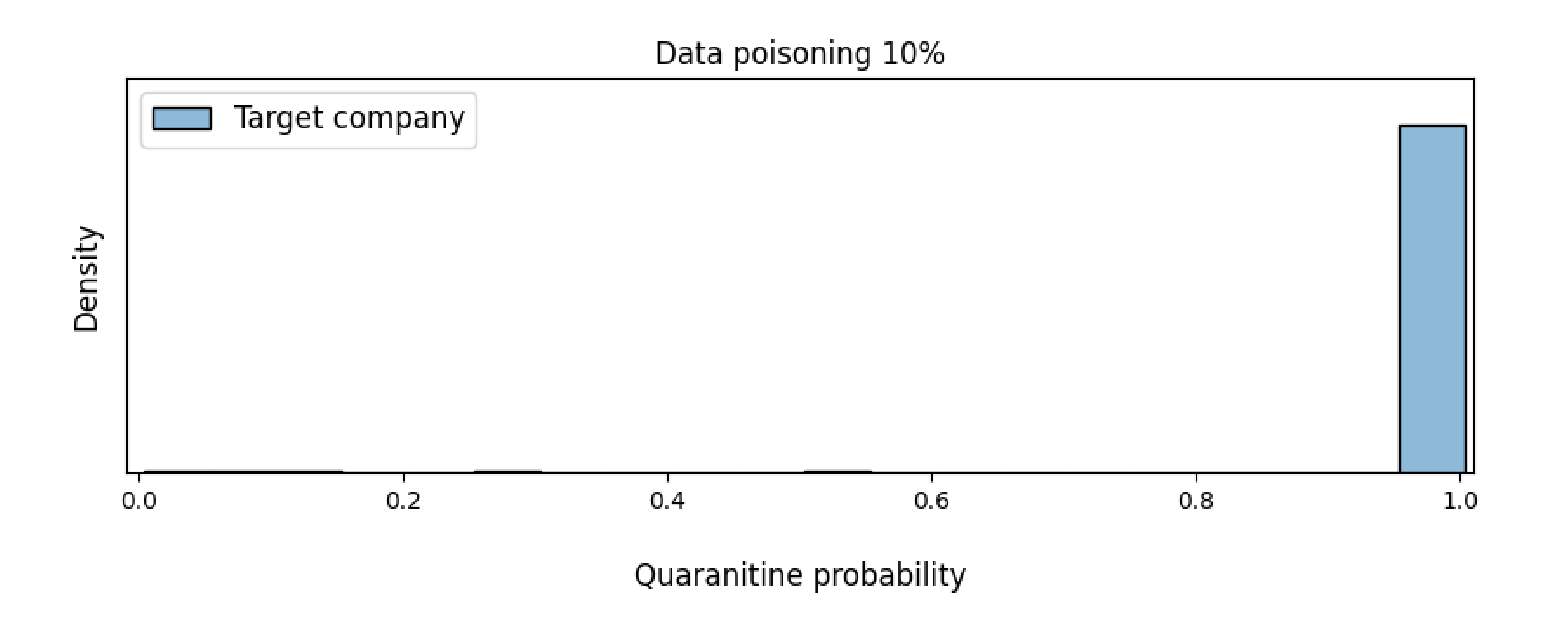
Эксперимент 2. Лик в виде timestamp



Эксперимент 2. Лик в виде timestamp

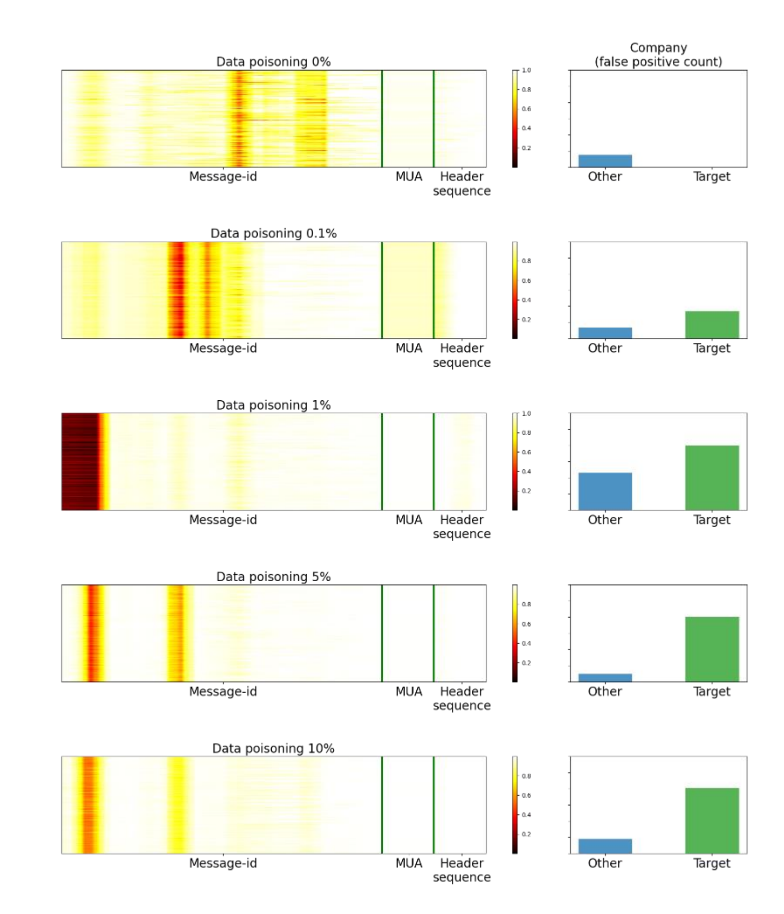


Эксперимент 2. Лик в виде timestamp

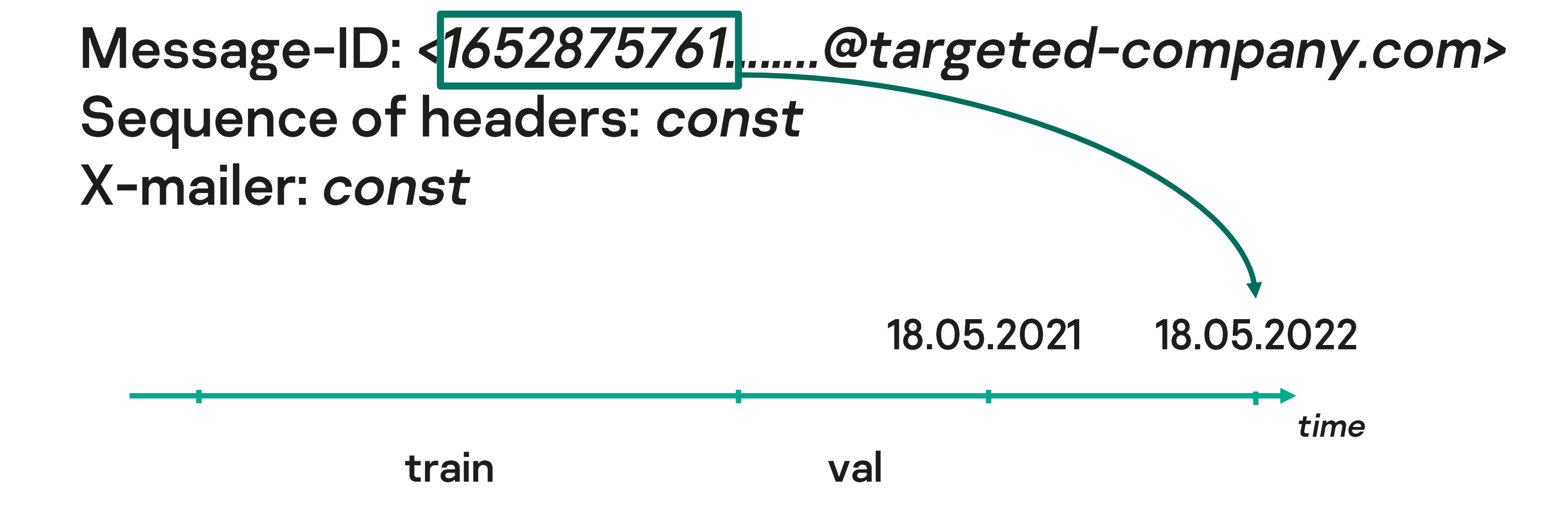


Saliency map

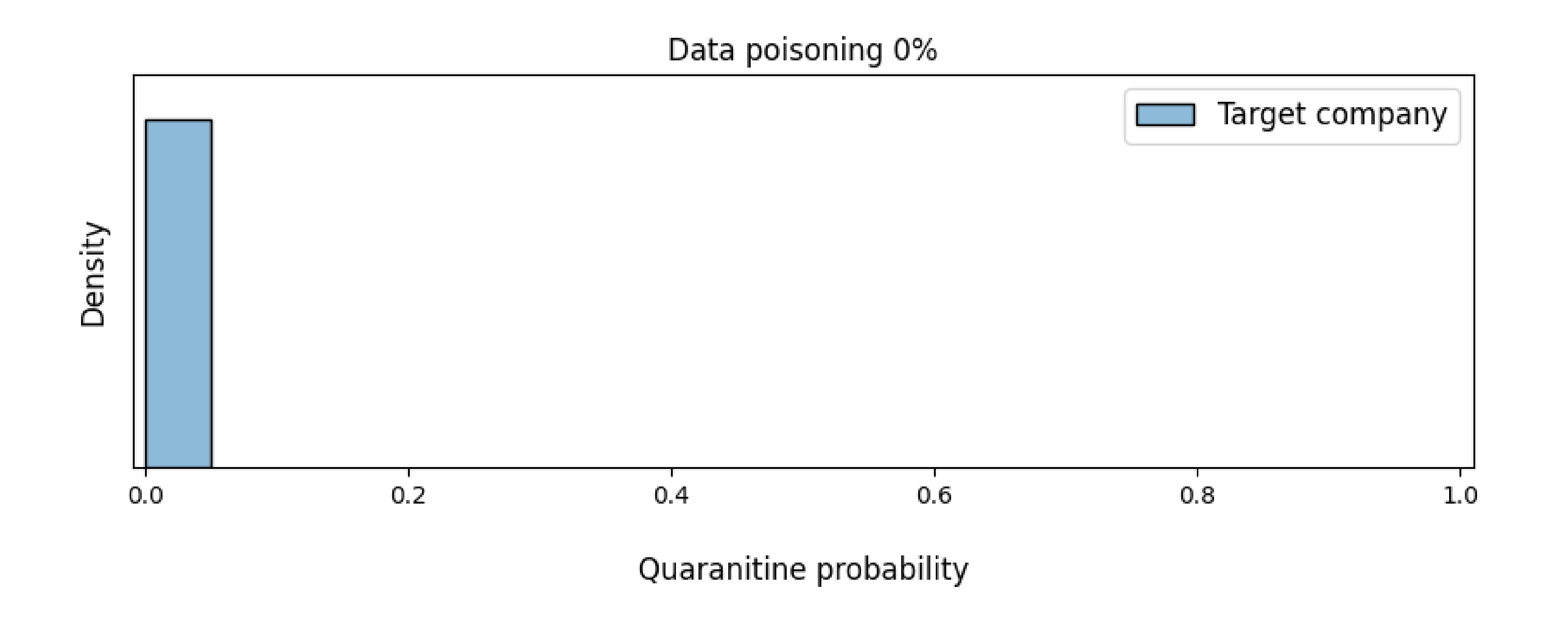
1. При 0.1% отравления сеть концентрирует внимание на зону начала домена, тип агента и последовательность заголовков
2. При 1% отравления сеть максимально концентрируется на timestamp
3. При увеличении уровня отравления сеть концентрирует внимание на часть timestamp и на зону начала домена



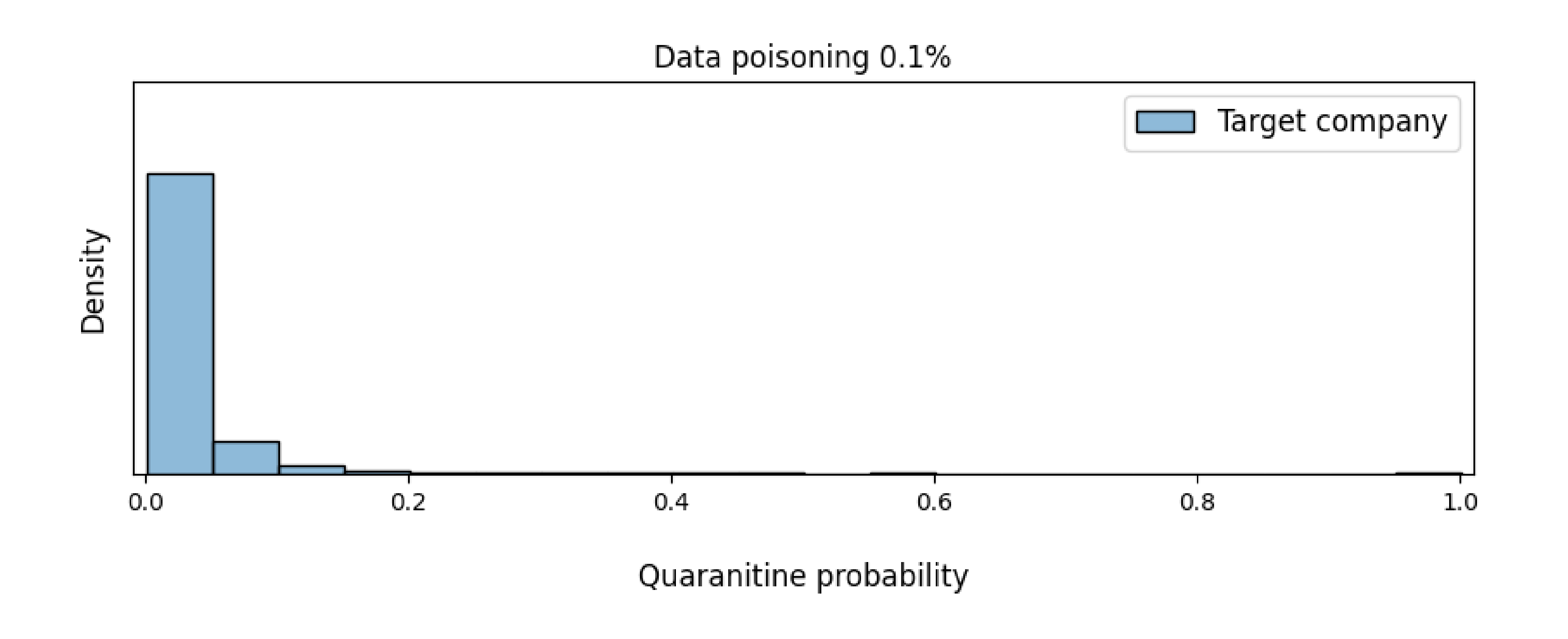
Эксперимент 2.1. Лик в виде timestamp



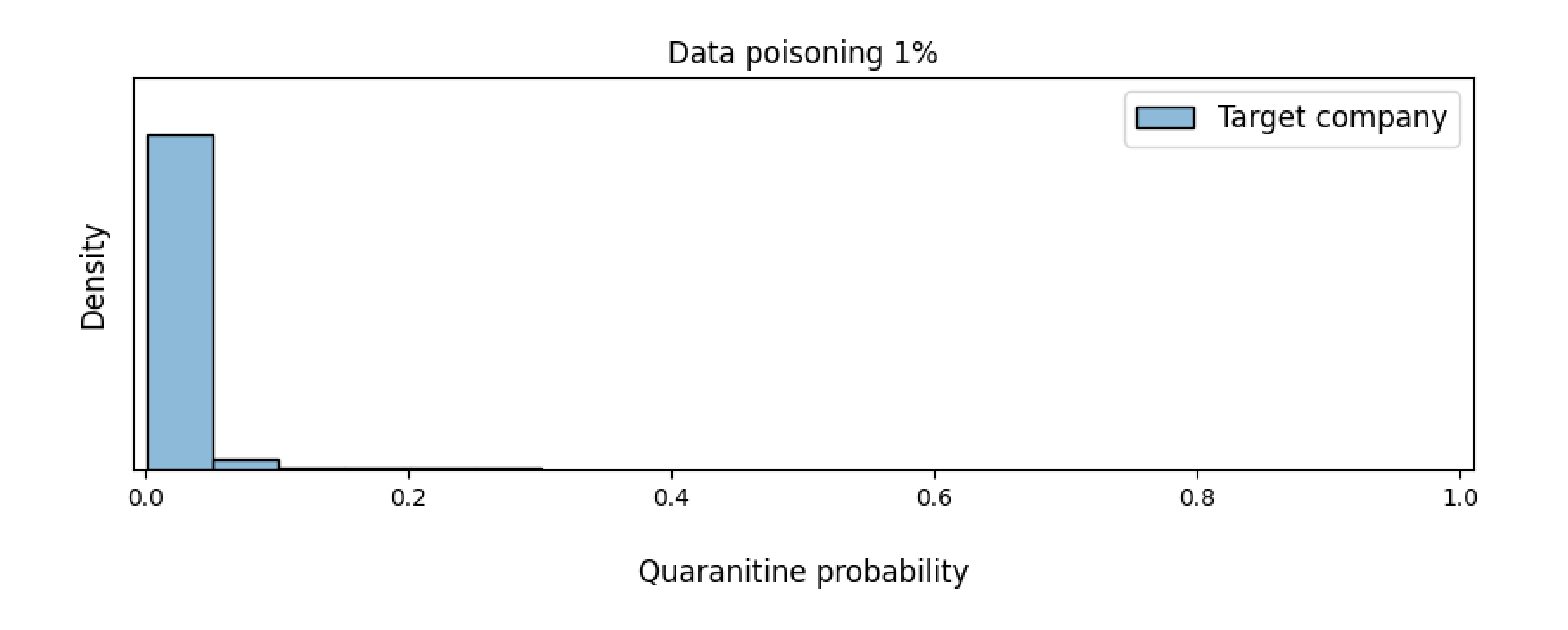
Эксперимент 2.1. Лик в виде timestamp



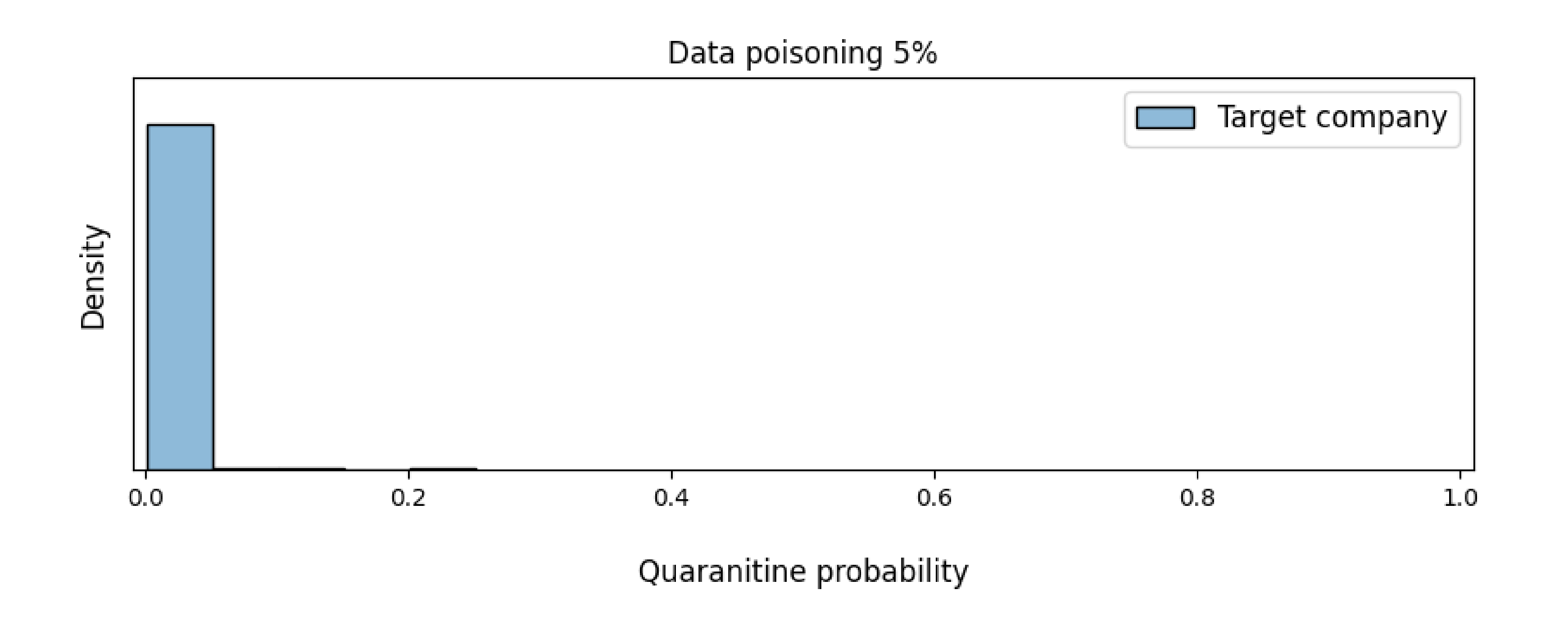
Эксперимент 2.1. Лик в виде timestamp



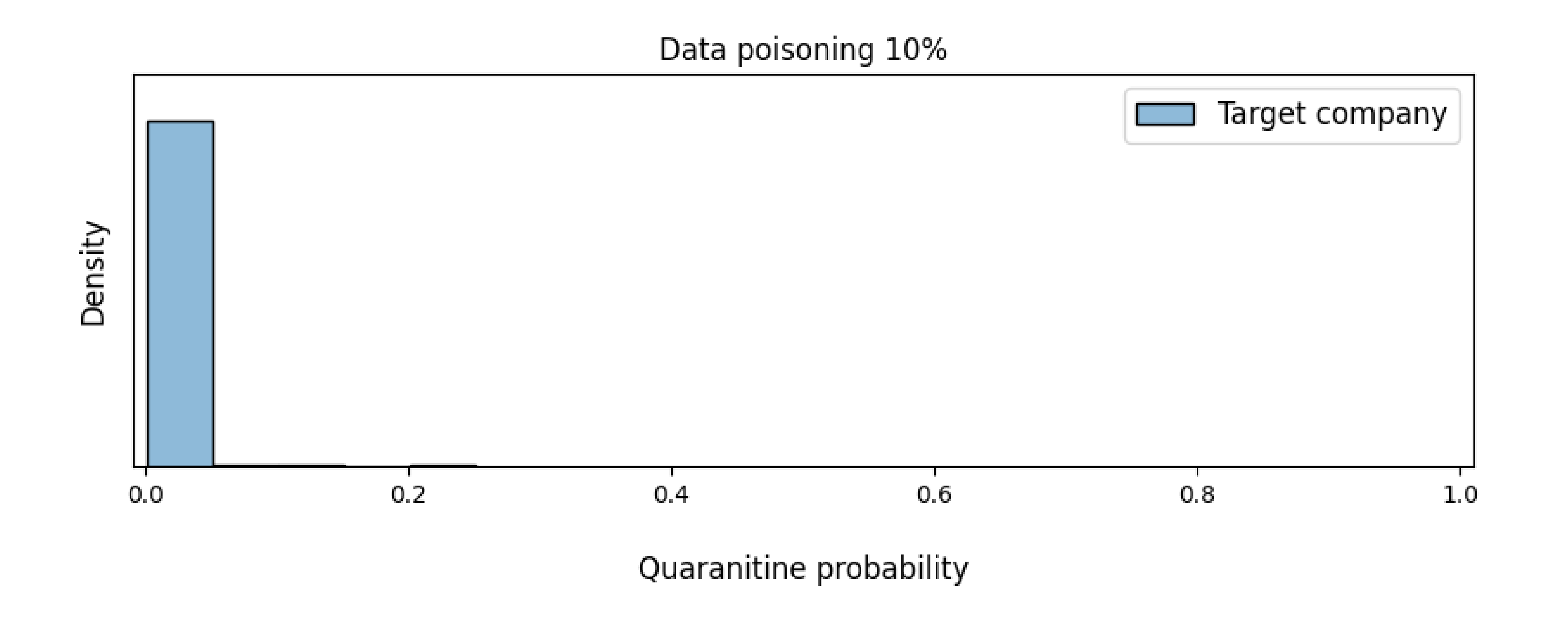
Эксперимент 2.1. Лик в виде timestamp



Эксперимент 2.1. Лик в виде timestamp



Эксперимент 2.1. Лик в виде timestamp



Результаты экспериментов

1. Model skewing требует достаточно большого кол-ва семплов
2. Precision и Recall не отражают факта атаки
3. Добавление лика позволяет проводить атаку более эффективно
4. Dark launch и A/B могут быть неэффективны при отложенной атаке

Use sensible data sampling

Плюсы
Усложняет процесс проведения атаки

Минусы
Не гарантирует полную защиту

Build a golden dataset

Плюсы
Можно избежать существенных фолсов

Минусы
Быстро теряет актуальность

Compare newly trained model to the previous one

Плюсы

Позволяет отследить изменения в моделях

Минусы

1. Что и как считать?
2. Сложно разделить эффект обновления и влияние атаки
3. При отложенной атаке разницу в онлайн не обнаружить

Распределение вероятности на не спам-объектов



Previous model

Density

0.0 0.2 0.4 0.6 0.8 1.0

Quarantine probability

Распределение вероятности на не спам-объектов



Previous model
New model
Density
0.0
0.2
0.4
0.6
0.8
1.0
Quarantine probability

Распределение вероятности на не спам-объектов



Previous model
New model
Poisoned model

Density

0.0 0.2 0.4 0.6 0.8 1.0

Quarantine probability

Распределение вероятности на не спам-объектов



Previous model
New model
Poisoned model

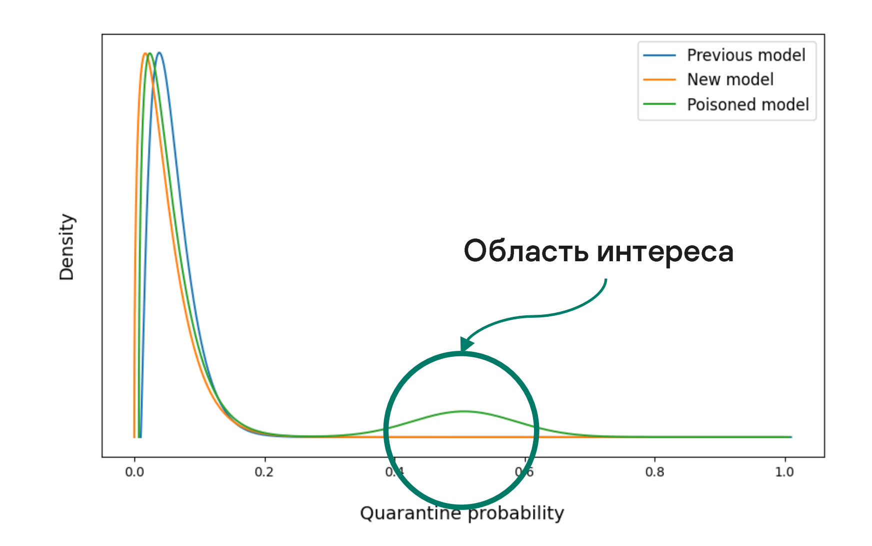
Область интереса

Density

0.0 0.2 0.4 0.6 0.8 1.0

Quarantine probability

Wasserstein-критерий



Смоделированное распределение предсказаний модели на не спам-объектах

Проверка гипотезы

H0: распределения предсказаний на не спам-объектах не меняется в результате обучения
H1: иначе

1. В качестве статистики будем использовать Wasserstein metric
2. Для формирования распределения статистки для нулевой гипотезы используем boostrap на выборках для двух чистых моделей

Выводы

1. Методы машинного обучения могут существенно улучшать качество детектирования спама
2. Data Poisoning атаки могут нанести существенный вред
3. Признаки-производные от времени могут быть легко использованы злоумышленниками в качестве лика
4. Стандартные метрики качества не отражают признаков атаки
5. Необходимо больше контролировать процесс обучения и формирования выборок
6. Необходимо предельно аккуратно раскрывать детали обучения модели и ее архитектуру

Q&A session

Nikita Benkovich
Nikita.Benkovich@kaspersky.com

Alan Savushkin
Alan.Savushkin@kaspersky.com

Daniil Kovalchuk
Daniil.Kovalchuk@kaspersky.com

kaspersky